Advances in the Theory of Nonlinear Analysis and its Applications 7 (2023) No. 4, 17-34.
https://doi.org/10.17762/atnaa.v7.i4.279
Available online at www.https://atnaea.org/
Research Article

Advances in the Theory of Nonlinear Analysis and its Applications

ISSN: 2587-2648 Peer-Reviewed Scientific Journal

Comparison between classical and robust estimation methods for regression model parameters in the case of incomplete data

HAYDER RAAID TALIB[a], HASSAN HOPOOP RAZAQ[b], SARAH ADEL MADHIOOM[c]

[a]*UNIVERSITY OF SUMER / COLLEGE OF ADMINISTRATION ECONOMICS.*
[b]*UNIVERSITY OF THI-QAR/ FACULTY OF ADMINISTRATION AND ECONOMICS.*
[c]*MIDDLE TECHNICAL UNIVERSITY SUWAIRA TECHNICAL INSTITUTE.*

Abstract

The problem of complete data is considered one of the most important problems that hinder statistical analysis. Therefore, this problem must be solved by finding sound solutions to it by using some methods that lead to accurate results or close to accuracy. The research aims to compare methods of statistical data analysis. Estimating incomplete data using the filling algorithms R-Estimators, L estimators, EM, and and the W method and the case of the normal distribution of data according to patterns, machines, and the method of losing these observations. A comparison was made between the robust methods and the maximum Likelihood method, and the efficiency of the above-mentioned methods was noted. The imitation method was used in addition to practice. the data on the function variables, which represent the demand for cash and its relationship with the gross domestic product, government consumer spending, and the consumer price index, were discussed for the period from 2000-2022.The process was to compare these methods, and the results were drawn to reach the experimental side , through the experimental and applied part, it is preferable to use robust methods over classical methods in the case of incomplete data.

Keywords: R-Estimators; L estimators; EM; W estimators; incomplete data; robust methods.
2010 MSC: 49J20, 49K20, 35Q35, 93C20.

Email addresses: haider.raid@uos.edu.iq (HAYDER RAAID TALIB), hasanhopop@utq.edu.iq (HASSAN HOPOOP RAZAQ), Sarah adel@mtu.edu.iq (SARAH ADEL MADHIOOM)

Received September 18, 2023; Accepted: November 18, 2023; Online: December 12, 2023.

1. Introduction

1.1 Incomplete data

The problem of missing data is one of the importantproblems that occur during statistical analysis. It is a realistic problem in various statistical studies, including surveys and field studies.

The problem of missing data causes inconsistency in estimating the model curve, especially when the number of missing values is large, which affects the characteristics of the sample drawn from the study population, which can lead to misleading conclusions in data analysis. Observational studies are more prone to the problem of missing data than other studies.

The data containing missing values affects the statistical analysis significantly. As it reduces the power of statistical inference. as it is known, the data used in statistical analyzes in general and the data used in regression models in particular are divided into three types: cross section data, time series data, and the hybrid between the two types, which is polling data.

The process of missing data in time series is usually more difficult than that in cross-sectional data for theoretical and philosophical considerations, which is usually determined by the pattern of data loss, and whether the estimation of these missing data or the estimation of the parameters of regression models in the presence of these missing data is fraught with many difficulties and complexities.

1.2 Patterns of Missing Data

The statistical methods for analyzing incomplete data depend on the type of data. Some of them are suitable for special patterns of incomplete data that can be arranged in a specific way, and these methods are usually clear in steps and easy to apply. As for the analysis methods suitable for the general pattern of data loss, they are more complex than the methods of special patterns, and this is what prompts many researchers to arrange They collected their data according to a regular pattern whenever possible, in order to avoid the use of computationally complex methods. Below are a number of incomplete data patterns represented by the values of the variables $X_1, X_2, \ldots\ldots, X_P$.

1) **The first pattern (Pattern of Univariate Missing Data)**
   This pattern represents the simplest case of incomplete data cases in which all variables are fully observed except for one variable that contains missing values in a section of its observations. This case is called Univariate Missing-Data. The following figure shows this pattern of incomplete data.

2) **The second pattern(Monotone or Nested Missing Data)**
   In this style, the data of the explanatory variables are arranged according to the number of missing values in an ascending or descending manner, so that the order of the fully observed variable is first, followed by the variable that includes the least number of missing values, and so on to the last variable that contains the largest number of missing values. Or we start with the variable that includes the largest number of missing values, and so on until you reach the last variable that is fully visible. This pattern is called the Monotone or Nested Missing-Data pattern, as in the figure below.
   In the event that the observations of two or more variables are equal, the variables are arranged according to their occurrence and are treated on the basis of being one sector consisting of a number of variables.

3) **The Third pattern (Missing Data with Unidentified Parameters)**
   It is the last pattern of special patterns, and this pattern is in the case of observations X_j, X_k that are not recorded in one observations, so any observation in the variable X_j corresponds to a missing observation in the variable X_k, and this case occurs when two samples are merged or synthesized and the following figure shows this pattern of incomplete data.

4) The fourth pattern (General Pattern)
The General Pattern, which is the pattern in which the data loss is random for any of the values of the studied variables. The following figure represents an example of missing data according to the general pattern.

1.3 The missing data mechanism

Statistical methods for analyzing incomplete data differ in their hypotheses about the mechanism that leads to data loss. Understanding this mechanism and determining its nature helps a lot in choosing the appropriate method for analysis, but is the entrance to diagnosing the method whose results are close to optimization for the data studied.

It is not possible to summarize the relationship of data loss for a particular variable with the values of the variable itself or with the values of other variables as follows:

1. The missing values of X_j are independent of the values of other variables and of the missing values themselves.
2. The missing value of X_j depends on the missing value itself.
3. The dependence of the missing values of X_j on the values of other variables in the sample.

In the case of regression analysis, the missing value may depend on the values of other explanatory variables, i.e $X_1, X_2, X_3, \ldots..X_{j-1}, X_{j+1}, \ldots..X_k$ or the values of the explanatory variables and the value of the response variable Y.

Thus, the mechanism can be divided as follows:

1) Missing Complete At Random (MCAR).
2) Missing At Random (MAR).
3) Missing Not At Random (Not MAR).

In order to clarify the mechanism or how it occurs, if the cause of the loss is independent of the missing value itself and the values of the other variables in the sample, then it can be said that the data is completely randomly lost, but if the cause of the loss is related to the values of other variables only and is independent of the lost value, in such a case it is Random data loss. Sometimes the reason for the loss is due to the missing value, i.e. the loss of data is not random. The loss mechanism can be expressed mathematically through its distribution proposed by Rubin, D. B. (1974) represented by the conditional distribution of (X/R) and with unknown parameters $\Psi, P(R/X, \Psi)$ as:

X: A matrix that represents the real data of $(n \times p)$ order.

R: a binary matrix that takes values (1,0) corresponding to the matrix X. Since:

$$r_{ij} = \begin{cases} 1, & \text{if xij was known} \\ \\ 0, & \text{if xij was missing} \end{cases}$$

The matrix R is called the Missing Data Indicator Matrix, so if:

$$P(\frac{R}{X}, \Psi) = P(\frac{R}{\Psi}), \qquad \forall X \tag{1}$$

The data is missing completely at random (MCAR) Missing Complete At Random, as the distribution does not depend on the missing X_{miss} values nor on the observed X_{obs} values.

$$P(R/X,\Psi) = P(R/X_{obs},\Psi), \qquad \forall X_{miss} \tag{2}$$

The data is miss at random (MAR) Missing At Random. It is clear from the above formula that the distribution depended on the observed values X_{obs} of the variables and did not depend on the missing values. As for the case in which the distribution depends on the missing values X_{miss}, according to the following formula:

$$P(R/X,\Psi) = P(R/X_{miss},\Psi), \qquad \forall X_{miss} \tag{3}$$

It can be said that the data is not lost randomly (Not MAR). When analyzing this type of data, the distribution of the loss mechanism must be taken into consideration. In the case of (MCAR) and (MAR), the mechanism of loss can be neglected.

The problem of incomplete data is the fourth problem in the problems of regression models. It is the problem of multicollinearity, the problem of autocorrelation, and the problem of heterogeneity of variances. There are a number of methods to address this problem, namely.

1.4 robust estimates

In view of the growing interest in the subject of robust estimations and as a result of previous statistical studies, many methods of robust estimation of location and measurement have been found and developed, which are the M method, the R method, the L method, and the W method. With regard to the L method, it is the estimators of linear combinations of ordered statistics, the R method is based on ranks, the M method is a modification of the principle of the greatest possibility method, and the W method is a generalization of the least squares estimation.

1) Linear estimators (L estimators)

The linear fit of the ordered statistic is called the L estimator. I assume that $x_{(1)} \le x_{(2)} \le ... \le x_{(n)}$ It is an ordered statistic for a sample of size n. I assume $a_1,\ldots,a_n$ real numbers, since that $0 \le a_i \le 1$; $i = 1,\ldots,n$ So that $\sum_{i=1}^{n} a_i = 1$ The general mathematical formula for the linear estimator for the location parameter is:

$$T = \sum_{i=1}^{n} a_i x_{(i)} \tag{4}$$

As the values of the weights $(a_i, i = 1,2,\ldots,n)$ are at their highest limit in the center of the sample, and the values of those weights decrease to be at their minimum as the observation rank moves away from the center of the sample, i.e. $a_1 = a_{(n+1-i)}; i = 1,2,...,n/2$ For this reason, it provides us with unbiased estimates θ, and that the center of symmetry is understood on the basis that the sample mean and the median have similar weights, and that the L values are calculated in the event that the random variable $X_{(i)}$ is assumed to have a symmetric distribution around the location and measurement parameter. Here are some robust estimators for the location parameter.

- **The Sample Median (MED)**

The sample mean is linearly estimated according to $T = \sum_{i=1}^{n} a_i x_{(i)}$ When n is an odd number, the weights are:

$$a_i = \begin{cases} 1, & i = (n+1)/2 \\ \\ 0, & \text{Otherwise} \end{cases} \tag{5}$$

When n is an even number, the weights are:

$$a_i = \begin{cases} 1/2, & i = n/2 \\ \\ 1/2, & i = (n+1)/2 \\ \\ 0, & \text{Otherwise} \end{cases} \tag{6}$$

- **Median Absolute Deviation (MAD)**

$MAD = b_{med_i}|x_i - med_j x_j|; (j = 1, 2, ..., n)$

This formula was introduced by Hampel (1974), because med means the median calculated by the previous formula, and the value of $\{b = [1/\Phi^{-1}(3/4)] \cong 1.4826\}$ This is so that the MAD statistics are consistent when the observations follow the normal distribution, and that $[b = 2.0781, 1, 1.4826]$ is according to the assumed distribution (normal, Cauchy, exponential), respectively. This statistic is important, as it can be considered as a robust initial value, which enables us to compute estimators with more efficiency and immunity. The MAD has a breakdown point of 50%, which makes it a good estimator that is resistant to anomalies.

- **Trimmed Mean Tr(a1,a2)**

It was proposed by Deniell, and the idea of this estimator depends on amputating part of the two ends of the sample or from one of them and then calculating the arithmetic mean of the amputated sample Tr(b1,b2). It can be expressed in the form:

$$T = \sum_{i=1}^{n} a_i x_{(i)} \tag{7}$$

The weights take the following form:

$$a_i = \begin{cases} 1/(n-g); & g_1 + 1 \leq i \leq n - g_2 \\ \\ 0, & \text{Otherwise} \end{cases} \tag{8}$$

as b1 and b2 are the amputation ratios from the left and right ends of the sample, respectively; $(0 \leq b_1 + b_2 < 1)$ and that $(i = 1, 2)$; $(0 \leq (g_1 + g_2) < n/2) : g_i = [b_i \times n]$.

In the case of symmetric amputation from both ends of the sample, $(g_1 = g_2)$, and in the case of asymmetric amputation, then $(g_1 \neq g_2)$; And in the case of amputation from the right side, the $(g_1 = 0, g_2 > 0)$ The opposite is true in the case of a left-sided amputation.

There is another formula for (straightened arithmetic mean), where the weighted average of α for $T(\alpha)$:

$$T(\alpha) = \frac{1}{n(1-2\alpha)}\left\{(1-r)\left[x_{(g+1)} + x_{(n-g)}\right] + \sum_{i=g+2}^{n-g-1} x_{(i)}\right\} \tag{9}$$

For the weights $a_i, i = 1, 2, \ldots, n$:

$$a_i = \begin{cases} 0, & \text{if} \quad \text{i} \leq \text{g} \quad \text{or} \quad \text{i} \geq \text{n-g+1} \\ \frac{(1-r)}{n(1-2\alpha)}, & \text{if} \quad \text{i=g+1} \quad \text{or} \quad \text{i=n-g} \\ \frac{1}{n(1-2\alpha)}, & \text{if} \quad \text{g+2} \quad \leq i \leq \quad \text{n-g-1} \end{cases} \tag{10}$$

where $g = [\alpha n], \quad r = \alpha n - g$

- **Adoptive Trimmed Mean**

Hogg proposed a method to adapt the trimmed arithmetic mean in the case of symmetrical amputation on both sides by finding the value of $[\alpha n] = r$ which represents the integer part of a real number αn as that $(0 < \alpha < 1)$ by specifying a value α appropriateness. The determination of a value α depends on the strength of the tail of the symmetrical distribution whose position parameter is to be estimated, whose estimate is given by the formula:

$$Q_n = [\bar{U}_n(0.2) - \bar{L}_n(0.2)]/[\bar{U}_n(0.5) - \bar{L}_n(0.5)]$$

as it represents each of $\bar{L}_n(\alpha)$ and $\bar{U}_n(\alpha)$ an average $[\alpha n]$ From the observations of the lower and higher ranked statistics. Therefore, Hogg's choice of value α is defined as:

$$a_i = \begin{cases} \frac{1}{8} & Q_n < 1.81 \\ \frac{1}{4}, & 1.81 < Q_n < 1.87 \\ \frac{3}{8}, & 1.87 < Q_n \end{cases} \tag{11}$$

2) R-Estimators

Its name comes from the fact that it is derived from rank tests. The origin of this idea is due to an idea presented by Hodges and Lehmann, who used the following Walsh rates:

$W_{ij} = (x_i + x_j)/2; \quad (i = 1,2,...,n); \quad (j = 1,2,...,n)$

The previous rates are used to find the R estimators for the location parameter according to the following formula:

$$T_R = med\left\{\frac{x_i + x_j}{2}; i < j\right\} \tag{12}$$

As this formula is the most abbreviated in terms of computational steps compared to the formula in it $(i < j)$, Or in the long form for all pairs of values (i, j). As for the R estimators of the measurement parameter, Rousseeuw and Croux suggested the following statistic:

$$S_R = Cmed_i\{med_j \mid x_i - x_j \mid\} \tag{13}$$

Since for every i the median is calculated to $\{\mid x_i - x_j \mid; j = 1,2,...,n\}$ This results in n numbers that give a final estimate of SR, and the objective of the constant value c is to make the estimate consistent with the assumed distribution, as (c = 1.6982 , 0.7071 , 1.1926) according to the assumed distribution (normal, Cauchy, exponential) respectively. When calculating the above formula, the medi is the ordered statistic you ordered $\{h = [(n+1)/2]\}$, The inner median, medj, is the ordered statistic you ordered $[h = (n/2) + 1]$ The previous two researchers also suggested, in the same research, another estimator of the type of R estimators, in order to estimate the measurement parameter, which is the following estimator:

$$Q_R = d\{x_i - x_j; i < j\}_{(k)} \tag{14}$$

As d: represents a constant amount that aims to make the estimate consistent according to the assumed distribution, so it takes the values (3.476, 1.207, 1.0483) according to the assumed distribution (normal, Cauchy, exponential), respectively. and that :

$$k = \binom{h}{2} \cong \binom{n}{2}/4; \quad h = [n/2] + 1 \tag{15}$$

3) Expectation-Maximization EM-Algorithm

Expectation-Maximization Algorithm (EM) is a method to find estimates of the greatest possibility of model parameters, and means achieving the expectation of the greatest limit. It is an iterative algorithm for estimating ML for the case of Incomplete Data, and it is used in the case when it is not possible to find the second derivative. The beginning of the emergence of the algorithm was in 1977 by researchers Rubin, Laird & Dempster, and they were the first to call it the EM algorithm (Expectation Maximization Algorithm). In order to clarify the steps of this method under the assumption of ignoring the MAR loss mechanism and that the parameters Ψ and θ are distinct, the EM algorithm is used to clarify the reliability between the missing data X_{miss}. and the unknown parameter θ, since θ is superimposed as an initial value to find estimated values for X_{miss}. , where parameter θ is estimated from the observed X_{obs}. data. And then using the estimates of θ to estimate the incomplete data X_{miss}.

And that the vector distribution of the variables X is:

$$f\left(X_{obs}., \frac{X_{miss}}{\theta}\right) = f\left(\frac{X_{obs}.}{\theta}\right) f\left(\frac{X_{miss}.}{X_{obs}}., \theta\right) \tag{16}$$

From the above formula, we notice that any term in this formula is only a function of the parameter θ, then:

$$\ell(\theta/X_{obs}.,X_{miss}.) = \ell(\theta/X_{obs}.) + logf(X_{miss}./X_{obs}.,\theta) + c \tag{17}$$

$$\ell(\theta/X_{obs}.,X_{miss}.) = logf(X_{obs}./X_{miss}.,\theta)$$

It denotes the logarithm of the log-likelihood function of the variables vector X.

$$\ell\left(\frac{\theta}{X_{obs}}.\right) = logf\left(\frac{\theta}{X_{obs}}.\right) \tag{18}$$

It is the logarithm of the log-likelihood function for the observed data X_{obs}.
C: an arbitrary constant.

$f(X_{miss}./X_{obs}.,\theta)$:It is the prediction distribution of the missing data with the parameter θ constant, and this predictive distribution shows the interdependence between the parameter θ and the missing values X_{miss}.

As it is clear from the above formula that the right side cannot be calculated due to the existence of $X_{miss}.$, and to solve it, both sides of the above formula are multiplied by the Prediction Distribution $(X_{miss}./X_{obs}.,\theta^{(t)})$ since $\theta^{(t)}$ is a preliminary estimate for the unknown parameter θ and by taking the integral with respect to X_{miss}. We get:

$$Q\left(\frac{\theta}{\theta^{(t)}}\right) = \ell\left(\frac{\theta}{X_{obs}}.\right) + H\left(\frac{\theta}{\theta^{(t)}}\right) + C \tag{19}$$

$$Q(\theta/\theta^{(t)}) = \int \ell(\theta/X_{obs}.,X_{miss}) f(X_{miss}./X_{obs}.,\theta(t)) \quad dX_{miss}. \tag{20}$$

$$H(\theta/\theta^{(t)}) = \int Log(X_{miss}./X_{obs}.,\theta) f(X_{miss}./X_{obs}.,\theta^{(t)}) \quad dX_{miss}. \tag{21}$$

And if we assume that $\theta^{(t+1)}$ is the value of θ that maximizes $Q(\theta/\theta^{(t)})$, then $\theta^{(t+1)}$ is the best estimate for $\theta^{(t)}$. To clarify this, then:

$$\ell(\theta^{(t+1)}/X_{obs}) \geq \ell(\theta^{(t)}/X_{obs}) \tag{22}$$

Therefore, the above formula can be explained as follows:

$$\ell(\theta^{(t+1)}/X_{obs}) - \ell(\theta^{(t)}/X_{obs}) = Q(\theta^{(t+1)}/\theta^{(t)}) - Q(\theta^{(t)}/\theta^{(t)}) + H(\theta^{(t)}/\theta^{(t)}) - H(\theta^{(t+1)}/\theta^{(t)})$$

From the above formula, we can see that the expression $Q(\theta^{(t+1)}/\theta^{(t)}) - Q(\theta^{(t)}/\theta^{(t)}) -$ is a non-negative quantity because the parameter $\theta^{(t+1)}$ It was chosen according to the following formula:

$$Q(\theta^{(t+1)}/\theta^{(t)}) \geq Q(\theta^{(t)}/\theta^{(t)}) \tag{23}$$

The quantity $H(\theta^{(t)}/\theta^{(t)}) - H(\theta^{(t+1)}/\theta^{(t)})$ can be proved to be a non-negative quantity using Jensen's inequality.

Below are the steps of the algorithm:

The first step (E-step) of the function $Q(\theta/\theta^{(t)})$ is done by integrating with x_{miss}. The logarithm of the possibilities of the function $\ell(\theta^{(t)}/x_{obs}., x_{miss}.)$ combined with the predictive distribution $(x_{miss}/x_{obs}., \theta)$.

The second step (maximization step M-step) for the parameter $\theta(t+1)$ is found by maximizing the function $Q(\theta/\theta^{(t)})$.

Convergence between iterations increases with a decrease in the percentage of missing observations.

There are obstacles in the case of applying the EM algorithm, which are:

I. In the case of missing large amounts of data, the application of the algorithm is very slow.

II. The difficulty of step M (step M) in a no closed form.

4) W-Estimation

The W estimator is a substitute for the M estimator, because (Tn) is the M estimator defined in the following form:

$$\sum_{i=1}^{n} \psi\left(\frac{x_i - T_n}{CS_n}\right) \tag{24}$$

When we define W according to the formula $UW(u) = \psi(u)$ and compensate for the value $\psi(w)$ by equalizing it we get:

$$\sum_{i=1}^{n} \left(\frac{x_i - T_n}{CS_n}\right) W\left(\frac{x_i - T_n}{CS_n}\right) = 0 \tag{25}$$

or by rearranging:

$$T = \frac{\sum_{i=1}^{n} x_i w[(x_i - T_n)/CS_n]}{\sum_{i=1}^{n} w[(x_i - T_n)/CS_n]} \tag{26}$$

Since Tn is the weighted average of x_i.

The algebraic solution to the above equation, as in the case of obtaining the average, is rarely available, so you solve a numerical solution. Assuming Tk is the estimate at k iterations. then:

$$u_i^{(k)} = \frac{x_i - T_n^{(k)}}{CS_n} \tag{27}$$

Then the iterative form becomes:

$$T_n^{(k+1)} = \frac{\sum_{i=1}^{n} x_i w[u_i^{(k)}]}{\sum_{i=1}^{n} w[u_i^{(k)}]} \tag{28}$$

The above equation is known as Iteratively Reweighed Least-Square (IRLS) with the weights $W_i^{(k)} = W[U_i^{(k)}]$.

By the method of ordinary least squares (OLS) we estimate T which in turn reduces the value $\sum_{i=1}^{n}(x_i - t)^2$ And it is in the form:

$$T = \frac{\sum_{i=1}^{n} x_i}{n} \tag{29}$$

also, the weighted least squares estimator Tw depends on the fixed weights Wi, and the choice of t that reduces the amount $\sum_{i=1}^{n} w_i(x_i - t)^2$ which is in the form:

$$T_w = \frac{\sum_{i=1}^{n} w_i(x_i - t)}{\sum_{i=1}^{n} w_i} \tag{30}$$

An iterative weighted least square estimation (IRLS) starts with an appropriate initial estimate of the W estimator for the location, iterates up a series of estimates, and stops when successive estimates converge to a certain level of accuracy.

It is noted that W is a generalization of the least squares estimate, just as the M estimate is a generalization of the greatest possible estimate. In some cases, the two estimates give the same results, and in general the two estimates are close.

1.5 Maximum likelihood estimates for incomplete data

Estimating the Maximum likelihood of complete data is not much different from the estimates of the greatest possibility in the case of a data loss problem, as here the possibility function is segmented, according to the nature and pattern of incomplete data, and after segmentation, the same method of derivation is followed as in the complete data, and then equal to zero to obtain Model parameters estimators.

It should be noted here that the greatest possibility function used in the analysis of incomplete data is determined on the basis of the data loss mechanism, which is the key to determining the analysis method. To clarify this, let us assume that:

Xobs.:The observed values actually represent the explanatory variable X.
Xmiss. The missing values represent the explanatory variable X.

To find the probability density function for the joint distribution (Xobs.,Xmiss.) it is as follows:

$$f\left(\frac{X}{\theta}\right) = f\left(X_{obs}\cdot, \frac{X_{miss.}}{\theta}\right) \tag{31}$$

To find the marginal probability density function for (Xobs.), this is done by integrating the function with respect to (Xmiss.) as follows:

$$f(X_{obs}/\theta) = \int f(X_{obs}\cdot, x_{miss}/\theta) dX_{miss} \tag{32}$$

You will notice that the function depends on the values of Xobs. Xmiss. is neglected. If it is used to estimate the distribution parameters, the loss mechanism is neglected. The estimate for the greatest possibility is done by maximizing the function L(θ/Xobs.) as:

$$L(\theta/X_{obs.}) \propto f(X_{obs.}/\theta)$$

$\propto$:indicates proportionality.
The above relationship is used if the missing values are Xmiss. Do not depend on Xmiss values. It can depend on the value of another variable, i.e.:

$$f\left(\frac{R}{x_{obs}}.,x_{miss}.,\Psi\right)=f\left(\frac{R}{x_{obs}}.,\Psi\right) \tag{33}$$

We can refer to the incomplete data pattern by the R matrix, which is the matrix of missing values and is called the Response Indicator Matrix.

It is a binary array whose elements take the value (1) if (Xij) is observation and (0) if (Xij)is missing. Rij can be treated as a random variable because the loss is random and the joint distribution of R, X is determined as follows:

$$f\left(X,\frac{R}{\theta},\Psi\right)=f\left(\frac{X}{\theta}\right)f\left(\frac{R}{X},\Psi\right) \tag{34}$$

$f(X/\theta)$: the represents the marginal distribution of X with unknown parameters that are θ.

$f(R/X,\Psi)$ the represents the conditional distribution of R/X .It is called the distribution of the loss mechanism with unknown parameters, Ψ.

The distribution of the observed data is generated through the integration of Xmiss. And as follows:

$$f(X_{obs}.,R/\theta,\theta)=\int f(X_{obs}.,X_{miss}./\theta)f(R/X_{obs}.,X_{miss}.,\Psi)dX_{miss} \tag{35}$$

$$L\left(\theta,\frac{\Psi}{X_{obs}}.,R\right)\propto f\left(X_{obs}.,\frac{R}{\theta}.,\Psi\right) \tag{36}$$

The above relationship is used if the missing values are related to the Xmiss. value. And it requires taking into account the distribution of the loss mechanism, that is, the inference depends on the function in the above formula, and accordingly, and from the previous formula, it is possible to obtain:

$$\begin{aligned} f(X_{obs}.,R/\theta,\theta) &= f(R/X_{obs}.,.,\Psi)\int f(X_{obs}.,X_{miss}.,\theta)dX_{miss} \\ &= f\left(\frac{R}{X_{obs}}.,\Psi\right)f\left(X_{obs}.\frac{\Psi}{\theta}\right) \end{aligned} \tag{37}$$

In most applications, the parameters Ψ, θ are distinct (distinct) and have a common domain for the parameters Ψ, θ that results from the domain of θ and the domain of Ψ. Accordingly, it is possible to infer the greatest possibility around θ through the function of possibility L(θ/Xobs.), and if the loss mechanism can be neglected, especially if the observed data is randomly lost MAR, there will be an increase in the estimated amount and the increase as a result of neglecting the loss mechanism.

Experimental part: There are several methods for generating random data that follow a normal distribution, including the Box-Muller method and the approximate asymptotical method, and the Box-Muller method is more popular for its ease of use and accuracy of its results, and for generating data distributed according to a multivariate normal distribution with a mean vector μ and a variance matrix $\sum, X\sim N(\mu,\sum)$, by following the following algorithm:

1. Generating zi values using the (Box Muller) method
$Z_i \sim N(0,1), \qquad i = 1,2,...,P$
$Z \sim MNp(0,1)$
Since Z is an $n \times p$ matrix.

2. Partition of the variance matrix $\sum$ using the (Cholesky decomposition) method to obtain $C.C^{'}C = \sum$ where C is a lower triangular matrix of $PtimesP$ order and is the square root of the matrix $\sum$. The segmentation process takes place as follows:

a) The di vector is calculated by the following formula:

$$d_i = a_{ij} - \sum_{k=1}^{j-1} \ell_{ik}^{(2)} d_k \tag{38}$$

$$\ell_{ij} = [a_{ij} - \sum_{k=1}^{j-1} \ell_{ik}\ell_{ik}d_k]d_j^{-1} \quad j,k = 1,2,...,P \tag{39}$$

b) The C matrix is extracted according to the following formula:

$$C = [LD^{\frac{1}{2}}] \quad \text{where} \quad D = \text{diag}(di) \quad \text{and} \quad L = [\ell_{ij}] \tag{40}$$

3. To calculate the matrix X whose data is distributed in a multivariate normal distribution with a mean vector μ and a covariance matrix and covariances is $\sum$ the following formula is used:

$$X^{'} = \mu + CZ^{'} \quad \text{and} \quad X \sim MN_p(\mu, \sum)$$

1.6 Generating Regression Model Data:

In addition to generating the X matrix, which represents the explanatory variables. The random error that follows a normal distribution is generated with a mean (0) and a variance ($\sigma^{(2)}$)

Then the values of the response variable are calculated from the following linear relationship:

$$Y = X\beta + e$$

The values of β are imposed according to the model to be studied and in a manner consistent with the nature of the studied phenomenon,depending on the theoretical background of the phenomenon.

The sample: Four sample sizes were used, small and medium, with two sizes, then the large ones, which are as follows (30, 50, 75, 125), and the error variance was in four values (1,1.5,3). We will address the problem of data loss and the case of the MAR loss mechanism with loss rates of 3%, 5%, and 10% and by repeating the experiment 1000 times in order to identify which algorithms are more efficient.

Analyzing the results of simulation experiments:

Table 1: MSE for the model to the estimation methods, loss ratios and sample sizes when $(e \sim N(0,1))$:

sample size	loss ratio	L Estimators	R Estimators	EM Estimators	W Estimators	MLE Estimators
30	%3	0.546	0.583	0.504	0.583	0.672
	%5	0.508	0.628	0.519	0.599	0.592
	%10	0.415	0.427	0.402	0.518	0.482
50	%3	0.448	0.588	0.410	0.474	0.469
	%5	0.519	0.528	0.507	0591	0.572
	%10	0.537	0.559	0.518	0.601	0.533
75	%3	0.617	0.603	0.663	0.672	0.626
	%5	0.623	0.602	0.658	0.680	0.662
	%10	0.628	0.606	0.698	0.627	0.663
125	%3	0.416	0.413	0.477	0.509	0.599
	%5	0.556	0.524	0.563	0.573	0.618
	%10	0.624	0.611	0.683	0.638	0.710

Table 2: MSE for the model to the estimation methods, loss ratios and sample sizes when $(e \sim N(0, 1.5))$:

sample size	loss ratio	L Estimators	R Estimators	EM Estimators	W Estimators	MLE Estimators
30	%3	0.603	0.637	0.638	0.662	0.704
	%5	0.636	0.658	0.669	0.683	0.713
	%10	0.698	0.720	0.734	0.722	0.793
50	%3	0.667	0.682	0.684	0.694	0.728
	%5	0.693	0.712	0.730	0.728	0.775
	%10	0.716	0.726	0.731	0.741	0.766
75	%3	0.537	0.499	0.527	0.563	0.583
	%5	0.549	0.503	0.527	0.559	0.606
	%10	0.628	0.610	0.621	0.623	0.722
125	%3	0.527	0.517	0.523	0.592	0.603
	%5	0.593	0.548	0.583	0.603	0.694
	%10	0.674	0.627	0.668	0.698	0.728

Table 3: MSE for the model to the estimation methods, loss ratios and sample sizes when $(e \sim N(0,3))$:

sample size	loss ratio	L Estimators	R Estimators	EM Estimators	W Estimators	MLE Estimators
30	%3	0.646	0.674	0.658	0.687	0.727
	%5	0.670	0.683	0.673	0.688	0.738
	%10	0.728	0.748	0.738	0.776	0.794
50	%3	0.743	0.798	0.793	0.802	0.839
	%5	0.793	0.798	0.795	0.837	0.839
	%10	0.836	0.869	0.863	0.884	0.936
75	%3	0.548	0.546	0.502	0.503	0.599
	%5	0.603	0.654	0.597	0.600	0.704
	%10	0.663	0.684	0.625	0.643	0.778
125	%3	0.485	0.504	0.447	0.477	0.589
	%5	0.575	0.592	0.558	0.561	0.638
	%10	0.667	0.694	0.620	0.654	0.772

The study included the case of the multiple linear regression model according to the following formula:

$$y_i = 0.67 + 0.646x_{i1} + 0.587x_{i2} + 0.669x_{i3} + e_i$$

Which suffers from a loss in the observations of its explanatory variables X's and under the assumption of a loss mechanism as follows:

Analysis of the results of simulation experiments of loss patterns:

First: In the case of the error variance, it is 1, From Tables (1) a, we note the following:

In the case of a sample size of 30 and 50 we notice that the EM method gave the best results for all loss ratios, then it was followed by the EM method and then the L method.

In the case of a sample size of 75 and 125 we notice that the R method gave the best results for all loss ratios, then it was followed by the R method and then the L method.

Second: In the case of the error variance, it is valued at 1.5, and from table (2), we notice the following:

In the case of a sample size of 30 and 50 we notice that the L method gave the best results for all loss ratios, then it was followed by the L method and then the R method.

In the case of a sample size of 75 and 125 we notice that the R method gave the best results for all loss ratios, then it was followed by the R method and then the EM method.

Third: In the case of the error variation being 3, and from Table (3), we note the following:

In the case of a sample size of 30 and 50 we notice that the L method gave the best results for all loss ratios, then it was followed by the L method and then the EM method.

In the case of a sample size of 75 and 125 we notice that the EM method gave the best results for all loss ratios, then it was followed by the EM method and then the W method.

Applied part: In this aspect, data on the function variables, which represent the demand for cash and its relationship to the gross domestic product, government consumer spending, and the consumer price index, were dealt with for the period from 2000-2022, which were taken from publications in the Central Bureau of Statistics and Information Technology of the Ministry of Planning and Development Cooperation:

Table 4: MSE values for the model, in the case of MAR loss mechanism and for loss modes classified according to estimation methods and loss ratios.

MSE / Methods	L Estimators	R Estimators	EM Estimators	W Estimators	MLE Estimators
3%	1.05634	1.17689	1.11652	1.20001	1.89332
5%	1.15164	1.02657	1.13527	1.17637	1.93241
10%	1.54718	1.27642	1.13467	1.38188	2.16552

Conclusions:

- When taking a loss rate of 3%, the L method was the best compared to the rest of the methods, and its result was better than the EM result, which ranked second if the sample observations were subject to loss. Then the R method ranked third, followed by the W method, ranked fourth, and the MLE method ranked last based on the MSE for each method.
- When taking a loss rate of 5%, the R method was the best compared to the rest of the methods, and its result was better than the EM result, which ranked second if the sample observations were subject to loss. Then the L method ranked third, followed by the W method, ranked fourth, and the MLE method ranked last based on the MSE for each method.
- When taking a loss rate of 10%, the EM method was the best compared to the rest of the methods, and its result was better than the R result, which ranked second if the sample observations were subject to loss. Then the W method ranked third, followed by the L method, ranked fourth, and the MLE method ranked last based on the MSE for each method.
- Through the experimental and applied part, it is preferable to use robust methods over classical methods in the case of incomplete data.

References

1. Basu, S. & Heitjan, D. F. (2006) "Distinguishing 'Missing at Random' & 'Missing completely at Random' The JASA, vol. 23, No. 4, p. 123- 132.
2. Beal, M.J., & Ghahramani, Z. (2016) "The variational Bayesian EM Algorithm for Incomplete Data: with Application to Scoring Graphical Model structures", www.stat.cmu.edu/fienberg/stat36-756-06.html.
3. Campbell, N. A (1999) "Robust procedures in Multivariate Analysis I: Robust covariance Estimation" Applied statistics, vol. 23, P. 213- 224.
4. David, T. O (2004) "Applied Robust statistics" Chapman & Hall, New York.
5. Dempster, A. P. & Larid, N. M. & Rubin, D. B. (1977) "Maximum Likelihood from Incomplete data via the EM Algorithm" J. Royal Statist. Soc. Ser. B, vol. 11, P. 1-22.
6. Honaker, J.; Joseph, A.; King, G. & Scheve, K. (2001) "Analyzing Incomplete political science data: An Alternative Algroithm for Multiple Imputation" American political science review, vol. 68, No. 2, P. 123-137.
7. Jank, W. (2006) "The EM Algorithm, it's stochastic Implementation and Global Optimization: some challenges and opportunities for OR", www.rhsmith.umd.edu/faculty/wjank/.../page1118.htm.
8. Jurckova, J. & Sen, P. K. (1996) "Robust statistical procedures; Asymptotic and Interrelations" John Wile & Sons, New York.
9. Lin, C. & Rubin, D. B. (2002) "Maximum Likelihood Estimation of Factor Analysis using the ECME Algorithm with complete and Incomplete Data", statistica sinica 8, p. 437-446.
10. Little, R. J. A. & Rubin, D. B. (2002) "Statistical Analysis with Missing Data", New York: John. Wiley.
11. Little, R. J. A. & Rubin, D. B (2003)" statistical analysis with missing data", 2nd ed., John Wile & sons, New York.
12. Maronna, R. A. (1976) "Robust M- Estimators of Multivariate location and scatter" The Annals of statistics, vol. 3, P. 33- 48.
13. Morrison, D. F. (1971) "Expectation and variances of Maximum likelihood Estimates of the Multivariate Normal Distribution parameters with missing data", JASA, vol. 53, P. 326- 339.
14. Peter, J. H (1981) "Robust statistic" John Wile & Sons, New York.
15. Roche. A. (2006) "EM Algorithm and variants: an informal tutorial", ftp://ftp.cea.fr/pub/dsv/madic/pubils/roche-em.
16. Rubin, D. B. (1974) "Characterizing the Estimation of parameters in Incomplete- Data problems" Journal of the American statistical association.

17. Bickel,D.R.,(2002). "Robust and efficient estimation of the mode of continuous data "http://www.mcg.edu/research/biostat/b
18. Daniel P. Normolle .(2003)."An Algorithm for Robust nonlinear Analysis of Radioimmunoassay and Othre Bioassays "Ann Arbor,MI 48109-2029 , 313-936-1013.